International Journal of Database Theory and Application
Vol.10, No.2 (2017), pp.31-40
http://dx.doi.org/10.14257/ijdta.2017.10.2.03

The Application of TF-IDF with Time Factor in the Cluster of Micro-blog Theme

Song Yu[1], Yang Wang[1], Tianchi Mo[1], Mingyang Liu[1], Hui Liu[2] and Zhifang Liao[1,*]

[1]*School of Software Central South University Changsha,China*
[2]*PhD Associate Professor Department of Computer Science Missouri State University 901 S. National Avenue Springfield, MO 65897*
Email: zfliao@csu.edu.cn

Abstract

Time factor is of great significance for the topic clustering for Micro-blog. Usually, the topics discussed most frequently during a certain period may become the hot issues. Therefore, this article has successfully obtained the method of TF-IDF-TF by different division of periods and setting of different weights, then applied it to the ULPIR Micro-blog content corpus, with the hierarchical clustering method and k-means method being used to make statistic analysis. The result of the experiment shows that, compared with the traditional TF-IDF(term frequency- inverse document frequency), the TF-IDF-TF method could provide more accurate clustering result, especially for specific topics during the period when users play most frequently.

Keywords*: micro-blog topic clustering, TF-IDF, time factor*

1. Introduction

As a public social network platform, Micro-blog has playing an enormous role in our current society, and a growing number of netizens choose to send or receive new information through Micro-blog. Up to September ,2015,the number of active micro-blog users has reached 222 million. With analysis of main important ideas per month of the active micro-blog users and its clustering , the hot topics can be summarized each month. While the analysis for the emotion of these topics could help to know about people's attitude toward these topics, thus providing convenience for public opinion monitoring, opinion poll, scientific research *etc.*

Compared with long text within certain format criterion or content limitation. Micro-blog is a kind of relatively oral words which can be personally released in near real time, with its word length being 140 words [1]. Therefore, the clustering of Micro-blog is quite different from that of other ordinary texts. While the biggest difference lies in the time-validity of the former, however, the current method cannot take advantage of its time-validity and it result in the loss of detailed topics on Micro-blog in the sea of data. This article is based on the study for NLPIR Micro-blog content texts [2], and recalculation on the former TF-IDF method, taking advantage of the time-validity of micro-blog, thus improving its feature words sorting and ultimately optimizing the clustering effect.

After the study of the traditional TF-IDF in the Micro-blog topics clustering research, this article pointed out the TF-TDF method which take time factor into consideration by giving different weight to different periods (increasing the value of certain feature item during period when most active micro-blog users plays while decreasing the value of the feature item during rest time.) and doing micro-blog topics clustering research to verify its effect.

ISSN: 2005-4270 IJDTA

2. Related Work

TF-IDF has been pointed out by Jones [3] ,which is a frequently-used weighting technique, mainly used for information retrieval and data mining, It is actually a kind of statistical method to evaluate the importance of words for corpus file. Its calculation formula is shown as (1):

$$w(i,k) = tf(i,k) * \log(N / n_i + 0.01) \tag{1}$$

Of which, w(i,k) refers to the weight of feature words i in text k. However, if the text is too long, the weight of the feature terms would increase. In order to remove the influence of long texts to the weight of feature terms, it is of necessity to normalize the vector quantity of the feature terms, that is the TFC method, and its calculation method [4] is showed as the formula (2):

$$w(i,k) = \frac{tf(i,k) * \log(N / n_i + 0.01)}{\sqrt{\sum_{i \in k} (tf(i,k))^2 \cdot [\log(N / n_i + 0.01)]^2}} \tag{2}$$

In the traditional TF-IDF method, the calculation of IDF contains large numbers of figures with no special meaning. So literature[5] mentioned improvement solution of TF-IDF method which is based on the category description, (TF-IDF-CD) and it mainly improves the feature weighting formula, introduces intra-class and inter-class information and amend TF-IDF weighting factor, and reduces problems of few and scattered high dimension of feature space caused by traditional TF-IDF method. Literature [6] pointed out a method integrating information entropy with TF-IDF together, mainly solving the problem that the inverse Document Frequency may be too scattered in the data concentration, thus boost the distribution accuracy of feature terms. Literature [7] improved the TF-IDF-IG method and pointed out the TF-IDF-IG-E method, which assuming that the intra-class distribution entropy of feature terms is in accordance with the classified information provided. However, it does not consider about that of inter-class information. Literature [8] adopted solution by square the Inverse Word Frequency, (IWF) to reduce the dependency of IDF on term frequency. While for the micro-blog, the methods above never consider about the time factor, and as a result, ideal topic clustering effect has not been achieved. Therefore, this article comes up with the TF-IDF method integrating time factor.

According to the statistic released by the Sina Micro-blog Data Center, the Micro-blog Users Development Report from 2012 to 2015 shows the average daily frequency of writing micro-blog of the users , and the details is shown as the flowing Figure 1:

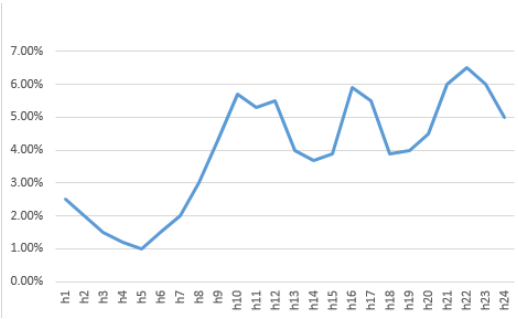

Figure 1. Percentage of Average 24hs Micro-blog Using in the Past 4 Years

From the graph, it can be seen that the value of micro-blog using percentage is closely related to the time. The value vary with the time and its variation is relatively large. Period with high value is called “active period” and that with low value is called “inactive

period". The importance of keywords extracted during active period is relatively larger than that from inactive period, and in this way, the real hot topic could be obtained.

3. TF-IDF-TF Method

3.1. Basic Idea

Micro-blog is a social network platform, and varieties of information can be created every seconds on it. Compared with other ordinary text clustering, the micro-blog information has its time-validity. Therefore, it is of great necessity to take time factor into account in the micro-blog topic clustering. During "active period", micro-blog users expressed more and feature words or terms created in such periods would be more important and should be paid higher weight, and lower weight for that in inactive periods. And in this way, the feature words or terms in active periods would be paid more attention and would not be replaced by the non-critical words, thus affecting the final feature words or terms ranking. According to graph 1, it can be found that the micro-blog users are quite independent on the micro-blog, and period from 8:00 am to 12:00 am, 15:30 pm to 17:30 pm, 21:00 pm to 24:00 pm is the so-called "active period", when users write most micro-blogs. And the rest of the time is inactive period. And the blog-using percentage in all periods are shown as Figure 2:

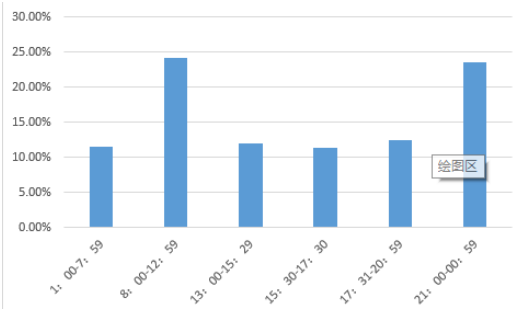

Figure 2. Blog-using Percentage in All Periods

So, in this article the whole day would be divided into 6 parts, including 3 active periods and 3 inactive periods. For different periods, different weight has been set thus to use TF-IDF method to extract the keywords. In the last step, the same keyword from different periods would be integrated and the final keywords ranking list would be available. Next, the method used in weight setting for different periods would be introduced.

3.2. Weight Calculation

According to the periods division mentioned in 3.1, there would exist different weight in different periods and the calculation of weight should meet the following demands:

(1) Weight value in active periods should be larger than that in inactive periods;

(2) The sum of weights in all periods should be 1.

Therefore, the calculation method of weights in all periods is shown as formula (3):

$$h_n = \frac{S_n / t_{ni}}{\sum_{1}^{n} S_n / t_{ni}} \tag{3}$$

Here, h_n refers to the weight in periods n, S_n refers to the percentage of micro-blog using percentage in period n, t_{ni} means that periods n contain i hours, $n \in (1,6)$.

In accordance with the calculation above, weights in all periods can be listed as Table 1:

Table 1. Weights in All Periods

No	Periods	weight
h1	1：00 - 7：59	0.06
h2	8：00 - 12：59	0.19
h3	13：00 - 15：29	0.15
h4	15：30 - 17：30	0.21
h5	17：31 - 20：59	0.16
h6	21：00 - 00：59	0.23

As the topics of micro-blog may last for some time, some feature words exacted during early period would re-appear during other periods, so it is necessary to combine and cope well with these words to figure out the real weight. And the final calculation method is shown as formula (4):

$$w(i,k)=\sum_{n}^{6} h_n \cdot tf(i,k) \cdot idf(i,k) \tag{4}$$

In this formula, $n \in (1,6)$, h_n refers to the weight during periods n, and the final feature words or terms ranking can be known.

3.3. Algorithm Steps

The new calculation TF-IDF (with time factor being considered) is described as follows:

Algorithm : TF-IDF-TF

Input : micro-blog texts prepared

Output: key feature words or letters

Step 1: divide the text data according to its post time

Step 2: use TF-IDF method to process the data after division

Step 3: multiply the feature words or terms in each period by the weight of corresponding period

Step 4: combine the same feature words or terms during different periods together according to formula (4), and leave the different ones alone.

Compared with traditional TF-IDF method, in steps 1, 3, and 4, time factor is additionally considered. And because of the features words has been sorted according to their posting time, the time complexity during division is set as O(T). While the computation of weight of the feature words or terms is related to its scale, the maximum length of feature words is set as N and the time complexity in step 3 and 4 set as O（N）. Therefore, the added time complexity created by new method should be O（T+N）.

4. Algorithm Implementation

The micro-blog topic clustering process consist of 4 parts including data preprocessing, feature words or terms matrix construction, topics clustering and clustering results evaluation.(shown as Figure 2)

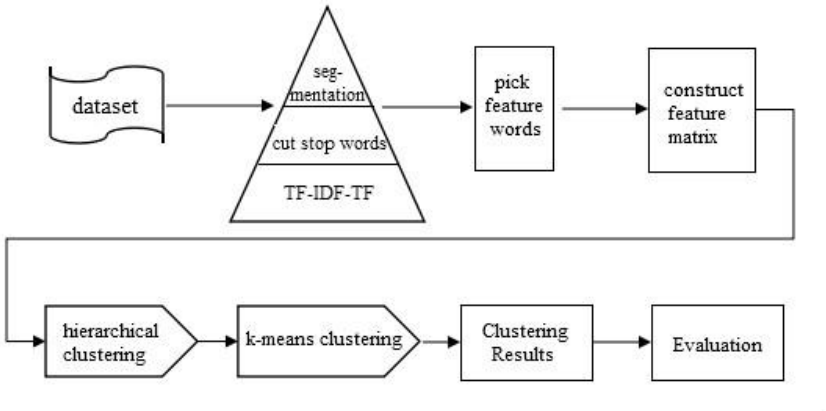

Figure 2. Micro-blog Topics Clustering Experiment Process

In this experiment, the programming language Python is adopted,with IDE being iPython notebook. The detailed experiment process is shown as follows:

(1) Data set preparation: the data comes from the NLPIR micro-blog content corpus, with its data saved as xml documents and each record contains 8 tags, including Article ID, text, comments quantity, text insertion time, source, users' ID, text post time and forwarding. While in this experiment, only 3 tags are needed: text code, text and text posting time. The first step should be the processing of these data and then form a trial data (I,T,C) for each record, (I refers to the No, T refers to the time and C refers to the text content).

(2) Preprocess: sort the data according to the time sequence, classify the words and delete those words out of use. Then divide the data into 6 sub data set in accordance with the method mentioned in the part 3 and utilize the TF-IDF method to extract the feature words and multiply it by its corresponding weight, and select the same feature words or terms from different periods together and figure out new feature terms and sort them according to their weight in descending order.

(3) Feature words or terms frequency matrix construction: as the micro-blog contains great information, without selecting and process of all the words and construct a word frequency matrix, its dimension would be very high and thin. So this experiment select the feature words obtained in the last step as the line of the word frequency matrix and the micro-blog user ID as the columns to record the numbers of each feature words' appearing. And the words frequency matrix would be shown as the following:

17	0	1	0	0	1	0	0	0	1	0	0	0	0	0
18	0	1	0	0	1	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	1	2	3	0	1	0	0	0	1
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	1	0	0
26	0	0	0	0	0	0	0	0	0	1	0	0	0	0

Figure 3. Matrix of Feature Words

(4) Topics clustering: For topic clustering, frequently used methods are LDA [9], hierarchy method [10], classification method [11]. For micro-blog topic clustering, it is not suitable to utilize LDA for it would tend to repeat the topic words for short texts. While simple as the hierarchy method is, it cannot modify the mistakes of combination or split, thus worsen the clustering effect. The k-means method is the most typical during all classification methods and it has the merits of high efficiency and it can cope with data of

great scale. But it is sensitive to outliers and it can only be used when the final number of cluster is confirmed. Therefore, this experiment chooses a kind of mixed clustering algorithm. The basic ideas are like this: clear the data whose record in feature words frequency remains 0, preliminary cluster the data with hierarchy clustering method, then get the number of clustering K and initial center, and input this in the k-means method and get the final clustering result. The detailed process are shown as the following:

The mixed clustering method
Input： feature terms frequency matrix；
$M = R \times C$, $R = \{id_1, id_2, ...id_n\}$, $C = \{t_1, t_2, ..t_k\}$
1. For each row $r \in R$:
2. If C_r all equal 0
3. delete r
4. Else for each $m \in M$:
5. use hierarchy clustering method
6. K←cluster number and initial center
7. use K-means clustering method
8. print result
Output: finally cluster number and extract topics

(5) Assessment: For the topics clustering, it is very difficult to make assessment for the clustering effect as it has no fixed indicators. Therefore, this experiment would make assessment for the clustering result of this method with the aid of manual annotation and other scholars' research. Literature [12] select some hot topics from the 1432 micro-blog records on Feb,1st ,2012,in NLPIR micro-blog content corpus, mainly including 4 hot topics: "Wu Ying, Death penalty, Fund raising", "Urban management officer, Law enforcement", "Corruption, Officials, Reform" and "Hong Kong, Mainland, Gravida, Tourist". Besides, through the artificial statistic of the micro-blog on that day, the topics discussed and comments quantity are listed as the Table 2:

Table 2. Topics Discussed on Feb. 1st, 2012

Topics	Quantity
Wu Ying, Death penalty, Fund raising	342
Urban management officer, Law enforcement, vendor	324
Corruption, Officials	322
Han Han, Fang Zhouzi	59
Hong Kong, Mainland, Gravida	29
Hong Kong, Mainland, Tourist	23
Hong Kong, iPhone, lottery	17
Others	328

The clustering result of the final micro-blog topics could be assessed by using the literature [12], and with the artificial statistics, the number of micro-blogs about each topic could be calculated accurately.

5. Experimental Results and Analysis

For better verifying the improvement of TF-IDF-TF algorithm, the result of traditional TF-IDF algorithm and that of TF-IDF-TF algorithm, has been compared. And all the data from NLPIR micro-blog content corpus on Feb. 1st, 2012 has been selected as the basic data set. And the result has been sorted in descending order by its weight of the feature words or terms. (shown as Table 3)

Table 3. Feature Words List Comparison between the 3 Algorithms

Algorithm	Feature terms(descending order)
TF-IDF	'Wu Ying', 'Urban management officer', 'Corruption', 'HongKong','Death penalty','micro-blog', 'China', 'Event', '10', 'Law enforcement', 'Attention', 'Society', 'Fund raising','Appeal to','2012','Vendor','Mainland'...
TF-IDF-CD	'Wu Ying','Urban management officer,' 'Corruption','Death penalty','HongKong',' Law enforcement', 'Fund raising', 'Vendor','mainland','public funds','case of Wu','Official'...
TF-IDF-TF	'Wu Ying','Urban management officer', 'HongKong', 'Corruption','Law enforcement','HanHan','Death penalty', 'Fang Zhouzi','mainland' ,'gravida', 'Fund raising','iPhone','lottery','tourist','Vendor',...

From the results ,it is not hard to find that the top of the feature terms list with 3 different algorithms are nearly the same, but the feature terms of TF-IDF algorithm contains lots of meaningless words and figure,(*e.g.*'event','10','society' *etc.*) while the feature terms of TF-IDF-CD algorithm has lost some important detailed words, (*e.g.*'iPhone','gravida','lottery' *etc.*) On the contrary, the feature terms list obtained by the TF-IDF-TF algorithm contains these detailed information and it indicates the importance of the time factor to the micro-blog.

For the 3 different feature terms above, 3 feature terms matrix has been created, with the mixed clustering algorithm, the TF-IDF algorithm deletes 382 data records and the final clustering result is shown as Figure 4:

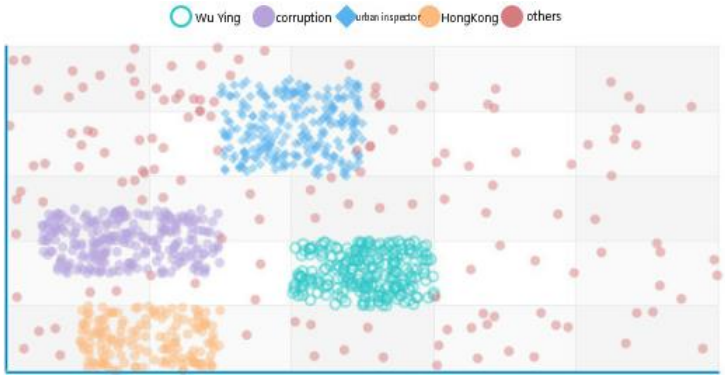

Figure 4. TF-IDF Clustering Result

With the mixed clustering algorithm, TF-IDF-CD algorithm delete 326 data records，the final clustering result is shown as Figure 5:

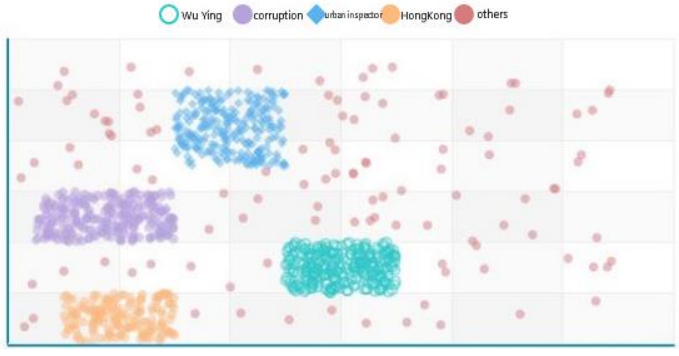

Figure 5. TF-IDF-CD Clustering Result

While the TF-IDF-TF algorithm delete 260 discrete data records and the final clustering result is shown as Figure 6:

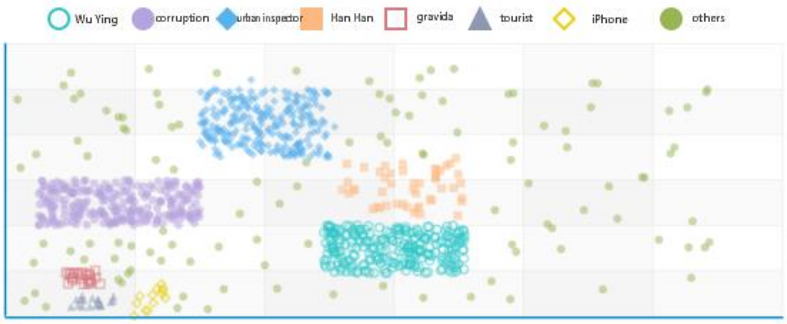

Figure 6. TF-IDF-TF Clustering Result

From the diagram, it can be noticed that the clustering result accuracy of the first and the second algorithm is far lower than that of the TF-IDF-TF algorithm as the former because lots of detailed topics has been lost. While through the evaluation mentioned in part 4 to test the clustering result, it can be concluded that the TF-IDF-TF algorithm can get better clustering effect as it can extract important words hidden in the sea of data, which has playing an vital role in the clustering process though its proportion is not very high. However, the traditional TF-IDF algorithm may extract some meaningless words which affect the accuracy of the topics clustering.

Besides, the feature terms extracted trough TF-IDF-TF algorithm is more typical than that extracted through the traditional TF-IDF algorithm, and it delete relatively smaller quantity of data during the discrete point- picking process, which affect the accuracy of data clustering effect. The comparison among the three algorithms can clearly shown as Figure 7 with the clustered topics on the abscissa and the clustering accuracy on the ordinate:

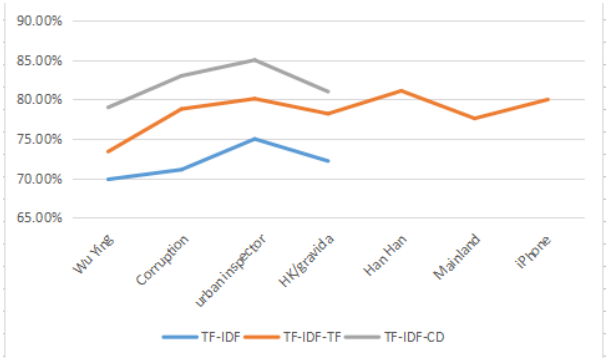

Figure 7. Comparison of the 3 Algorithms

For better verifying the reliability of the algorithms, this article also make a study on the micro-blog content separately on Feb. 2nd, 2012 and Feb. 8ths, 2012, and undertake clustering and analysis research, with the comparison of results shown as graph 8:

Figure 8. Result Comparison of 3 Algorithms

The experiment results shows that the TF-IDF-CD algorithm, with its clustering accuracy higher than the other method, is not dependent on the IDF algorithm. While the accuracy of clustering result with the improved TF-IDF-TF algorithm is 4.5% higher than that with TF-IDF method. However, the accuracy of results with both TF-IDF-CD algorithm and TF-IDF method are far from that with TF-IDF-TF algorithm. Through tests, the topics clustering accuracy with TF-IDF-TF algorithm increases by 31% than that with traditional TF-IDF method, especially for the detailed topics. Therefore, for the micro-blog topic clustering, the TF-IDF-TF method can greatly enhance the clustering precision for it never ignore any detailed information during any periods.

6. Conclusion

This article is based on the analysis for the Sina Micro-blog users and the statistics on the blog-releasing time of the past 4 years, as well as the division of the 24 hour in the whole day.

With higher weight value given to the user active period, this article points out the TF-IDF-TF method and it is found that such method could help to extract important feature words during different periods, in case that the detailed information is hidden and lost. Through experiment, it shows that such method can improve the accuracy of the clustering topics as well as that of the clustering effect or result. However, this method is highly independent on the feature words and the polysemy of features words has not been taken into consideration yet during clustering process. Therefore, it is necessary next to find solution to deal with the polysemy of features words and find better and more effective clustering method.

References

[1] P. Magistry, S. K. Hsieh and Y. Y. Chang, "Sentiment detection in micro-blogs using unsupervised chunk extraction", Lingua Sinica., vol. 2, **(2016)**, pp. 1-10.
[2] "NLPIR weibo corpus", natural language processing and information retrieval sharing platform(http://www. nlpir.org/).
[3] S. Luther, D. Berndt and D. Finch, "Using statistical text mining to supplement the development of an ontology", Journal of Biomedical Informatics, vol. 44, **(2011)**, pp. 86-93.
[4] X. Wang, L. Yang and D. Wang, "Improved TF-IDF Keyword Extraction Algorithm", Computer Science and Application, vol. 3, **(2013)**, pp. 64-68.
[5] X. D. Dong and W. S. Bo, "An Improved TF-IDF Feature Selection Based on Categorical Description", XIANDAI TUSHU QINGBAO JISHU, vol. 3, **(2015)**, pp. 39-48.
[6] J. Castro, R. M. Rodriguez and M. J. Barranco, "Weighting of Features in Content-Based Filtering with Entropy and Dependence Measures", International Journal of Computational Intelligence Systems, vol. 7, no. 1, **(2014)**, pp. 80-89.
[7] L. Ming, L. Rui and X. Liang, "TFIDF Algorithm Based on Information Gain and Information Entropy", Computer Engineering, vol. 38, no. 8, **(2012)**, pp. 37-40.
[8] T. Peng, L. Liu and W. Zuo, "PU text classification enhanced by term frequency–inverse document frequency-improved weighting", Concurrency & Computation Practice & Experience, vol. 26, no. 3, **(2014)**, pp. 728-741.
[9] R. Fu, B. Qin and T. Liu, "Open-categorical text classification based on multi-LDA models", Soft Computing, vol. 19, no. 1, **(2014)**, pp. 29-38.
[10] R. Cai, Z. Zhang and A. K. H. Tung, "A general framework of hierarchical clustering and its applications", Information Sciences, vol. 272, no. 3, **(2014)**, pp. 29-48.
[11] J. Aiwei, "Survey on partitional clustering algorithms", Electronics Design Engineering, vol. 22, no. 23, **(2014)**, pp. 38-41.
[12] Z. Qing and H. Ke,"News topic recognition from Chinese microblog base on word co-occurrence graph", CAAI Transactions on Intelligent Systems, vol. 7, no. 5, **(2012)**, pp. 444-449.

